

NCHS RDC Disclosure Manual: Rules for Researchers

Part 1: Confidentiality and the NCHS Research Data Center

Introduction

Working at the NCHS RDC allows you the opportunity to conduct research using confidential micro level data. This is a special privilege and it comes with special requirements. Your research will be subject to an in depth review at two stages: during the proposal process and during output review. Additionally, you will work in environments with additional rules and limitations to which you may not be accustomed. The intention of this document is to outline these constraints for you. The constraints stem from the legal requirements to meet the conditions of your appointment as a researcher at the RDC- particularly the requirement to maintain confidentiality of the underlying data sets to which you have been granted access.

Preventing Disclosure

The NCHS RDC strives to maintain confidentiality and prevent disclosure in three ways:

1. Research Proposal: The review committee (RDC Analyst, RDC Director, Confidentiality Officer, and representation from the data system(s)) carefully examine the variables requested and the plan of analysis. This approved document becomes a contract between the RDC and the researcher.
2. Confidentiality Orientation and Paperwork: researchers complete an online orientation about confidentiality and disclosure. Additionally, there are three legal documents that give researchers Designated Agent status and hold them accountable to the laws. These forms are specific to the project and must be completed each time you have a new approved proposal.
 - [Links to the confidentiality page on the website.](#)
3. Disclosure Review: All research output must be reviewed by an RDC Analyst before it can be released to the researcher.

It is important to note that the NCHS RDC's primary concern is with disclosure. We evaluate research proposals based on two major factors: 1) accordance with the NCHS mission to provide a public health benefit, and 2) risk of disclosure.

Confidentiality and the Law

There are two laws that govern the NCHS RDC: Section 308(d) of the Public Health Service Act and Confidential Information Protection and Statistical Efficiency Act (CIPSEA). In short, the Public Health Service Act asserts the importance of protecting confidentiality and that the only people who can access confidential data are NCHS staff and Designated Agents. Therefore, researchers wishing to access confidential data must become Designated Agents. CIPSEA stipulates the penalties for violating confidentiality as up to 5 years in prison and/or a \$250,000 fine. We are often asked about how the Freedom of Information Act (FOIA) applies: the short answer is that records collected under the Public Health Service Act are not subject to FOIA.

Section 308(d) of the Public Health Service Act 42 U.S.C. 242m(d)

"No information, if an establishment or person supplying the information or described in it is identifiable, obtained in the course of activities undertaken or supported under section 242b, 242k, or 242l of this title may be used for any purpose other than the purpose for which it was supplied unless such

NCHS RDC Disclosure Manual: Rules for Researchers

establishment or person has consented (as determined under regulations of the Secretary) to its use for such other purpose; and in the case of information obtained in the course of health statistical or epidemiological activities under section 242b or 242k of this title, such information may not be published or released in other form if the particular establishment or person supplying the information or described in it is identifiable unless such establishment or person has consented (as determined under regulations of the Secretary) to its publication or release in other form.”

<http://www.cdc.gov/nchs/data/misc/staffmanual2004.pdf>, page 1

Confidential Information Protection and Statistical Efficiency Act (CIPSEA)

“Whoever, being an officer, employee, or agent of an agency acquiring information for exclusively statistical purposes, . . . comes into possession of such information by reason of his or her being an officer, employee, or agent and, knowing that the disclosure of the specific information is prohibited under the provisions of this title, willfully discloses the information in any manner to a person or agency not entitled to receive it, shall be guilty of a class E felony and imprisoned for not more than 5 years, or fined not more than \$250,000, or both.”

<http://www.cdc.gov/nchs/data/misc/staffmanual2004.pdf>, page 2

Guidelines for NCHS RDC Researchers:

1. To conduct ONLY the analyses for which you have received approval. Failure to comply will result in cancellation of the research activity and potential disbarment from future research activities in the RDC.
2. To hold in strictest confidence the identification of any establishment or individual that may be inadvertently revealed in any documents or discussion, or analysis.
3. To immediately bring any inadvertent identification of any establishment or individual to the attention of RDC staff.
4. Do NOT use ANY technique in an attempt to learn the identity of any person, establishment, or sampling unit not identified on public use data files.
5. Do NOT remove data sets from the on site workstations.
6. Do NOT save output, files, or programs to transportable electronic media. RDC Analysts can copy approved output or programs to transportable media for you.
7. Do NOT remove any output or notes until they have been reviewed for disclosure risk by RDC staff.

NCHS RDC Disclosure Manual: Rules for Researchers

Part 2: The RDC Research Process

Proposal Process

*[*Link to Proposal Pages on the Website](#)*

Making Changes to Your Proposal

We understand that research evolves and may change from the day you submit the proposal to the end of your analysis. However, it is important that your RDC Analyst be made aware of changes throughout the process, as these changes may affect the disclosure risk. It is also important that the research proposal be updated to represent these changes.

- **New Researchers:** If the Primary Investigator or individuals in direct contact with the data change during the research project, two steps must be taken:
 1. Complete the Confidentiality Training and related paperwork.
 2. Update the proposal to reflect the change, including submission of the CV.
- **New Variables:** Addition of new variables related to the original research question must undergo a brief review.
 1. Update the proposal's data dictionaries to highlight the change and submit an explanatory email.
 2. If the data set has already been created, adding variables may result in additional costs.

Addition of variables unrelated to the original research will most likely require a new proposal and complete review.

- **New Methods or Types of Output:** If the analysis starts to take a different direction we must be informed as this may significantly change the disclosure risk.
 1. Discuss the changes with your RDC Analyst. He/she may decide additional review by the committee is required.
 2. Highlight these changes in your output explanations.

Providing the Public Use Data Set

Researchers are responsible for compiling the public use data set and providing any non-NCHS data that will be included in the merge. When you submit the public use data set to your RDC Analyst, please include a file that lists all of the variables and defines them (i.e. SAS users can submit a proc contents).

- **Original NCHS Variables** – must have the name they are given in the public use data set. If you want to rename the variables, you must include the original variable name in the variable description.
- **Derived Variables** – must be labeled with the variables from which they were derived and any arithmetic manipulation must be explained.
- **Public Use Mortality Variables** – if you are requesting access to the restricted Mortality files, you CANNOT include any public use mortality variables, or variables derived from the public use mortality data.
- If you have added variables that were not in the data dictionary in your approved proposal, please highlight these variables. They may require additional approval.
- Any attempt to include variables that may lead to re-identification of subjects/establishments is considered a disclosure violation and will result in the cessation of your project and possible legal actions.

NCHS RDC Disclosure Manual: Rules for Researchers

Using the Remote Access System

1. Remote access users can only submit statistical code related to the analysis plan outlined in the research proposal. If additional/different analyses are desired, they must first be discussed with your RDC Analyst.
2. Remote access rights are only granted to one person. Only that person can submit data requests and receive output from the remote access system (ANDRE).
3. Any attempt to circumvent ANDRE to obtain prohibited information that may result in a disclosure will result in an immediate suspension of the account and possible legal actions.
4. The remote access system (ANDRE) does NOT allow researchers to view any data at the individual record level.
5. Certain SAS procedures are not permitted on the remote access system. Remote access users, please read the complete list http://www.cdc.gov/nchs/data/r&d/SAS_Restrictions040809.pdf.
6. Output results that pose a disclosure risk will be suppressed. Additional cells or values may be suppressed to maintain confidentiality (complementary suppression).
7. Output will be returned in a text file via email.

NCHS RDC On Site Rules

When you come on site to NCHS RDCs or Census RDCs, you will experience restrictions that have been developed to decrease the likelihood of a disclosure.

1. RDC computers do not provide access to the local network, the mainframe, or the Internet.
2. RDC Analysts will merge the approved confidential variables to the public-use data and other data sources (if applicable).
3. If you have approval for multiple research projects, you will only be able to work on one project at a time. You may not have data sets for different projects opened simultaneously.
4. Cell phones, pagers, laptops, or other communication devices are not permitted in the RDC.
5. You are not allowed to bring items into the RDC that may enable you to identify individuals and/or establishments (e.g. a national hospital directory, information from voter registration).
6. All output and all notes written in the RDC are subject to disclosure review by RDC Analysts before removal from the RDC.
7. Output will be released in the default format produced by the statistical package (i.e. .lst files for SAS) or output can be released as rich text format (.rtf) files.

External Media in an NCHS RDC

You may bring statistical programming code with you to the NCHS RDC on external media (flash drive). The RDC Analyst will take the flash drive from you upon your arrival. The RDC Analyst will load the programs onto the workstation for you. The statistical code will be reviewed before use. If you would like to take the code with you at the end of the day, the RDC Analyst will need to review the code leaving the RDC before releasing it to you. The RDC Analyst will load it onto the flash drive provided by you.

- Do NOT attempt to put any content in your code that would facilitate re-identification of a subject/establishment. This is considered a disclosure violation and will result in the cessation of the project and possible legal actions.

NCHS RDC Disclosure Manual: Rules for Researchers

Scheduling Lab Time at an NCHS RDC

Both the Hyattsville and Atlanta RDCs are open Monday-Friday 9:00-5:00, except federal holidays. Time in the RDC may be scheduled by appointment. We recommend that you schedule at least 10 business days in advance. Priority will be given to researchers who schedule in advance and adhere to their scheduled visits. Contact your RDC Analyst to schedule time.

- There are additional procedures for international visitors that may require greater lead time. If you are an international researcher, please discuss this with your RDC Analyst early in the process, so as to avoid a delay.

Working at a Census RDC

While working at a Census RDC you are subject to all the same rules and restrictions as NCHS on-site users in addition to following rules:

1. To work in a Census RDC you must become an NCHS Designated agent and have Special Sworn Status. Census RDC Administrators will be available to assist you with obtaining Special Sworn Status.
2. You must request output review through the Census Tracking System. Your RDC Analyst will receive email notification of your request. This is not an automated process; please allow three weeks to receive your approved output.
3. Census RDC Administrators can not modify your dataset(s): you will need to contact your RDC Analyst.

Publications, Presentations, and Reports

Please provide all publications, presentations, and reports that refer to research conducted using the RDC. Please email them to rdca@cdc.gov and your RDC Analyst. Additionally, all publications, presentations, and reports must include the following disclaimer:

The findings and conclusions in this paper are those of the author(s) and do not necessarily represent the views of the Research Data Center, National Center for Health Statistics, Centers for Disease Control and Prevention.

Data Storage

Data and output will be stored in the RDC for two years. The researcher can initiate arrangements to store data longer. A reasonable effort (emailing and calling last known location) will be made to contact researchers before their data is disposed of.

NCHS RDC Disclosure Manual: Rules for Researchers

Part 3: Disclosure Review Policies and Procedures

Our disclosure review policies and procedures exist to protect the confidentiality of NCHS study participants. We strive not only to avoid disclosure of confidential information, but also the perception of disclosure. Perceived disclosure could have a negative effect on study participation and is seen as detrimental to studying the health of the population.

Data Set Policies

1. Absolutely NO individual level data will leave the RDC facilities.
2. Confidential data will be merged to public use data by an RDC Analyst.
3. Researchers may derive variables from the data set provided to create new variables, but they cannot introduce new data using their statistical code.

General Output Policies

1. Absolutely NO output will leave the RDC facilities without first being reviewed by an RDC Analyst. Typically output is returned via email.
 - NCHS RDC Users: Output can be at the conclusion of the research project OR once per week. It will be returned within 5 business days unless a more extensive review is required.
 - Census RDC Users: Output will be reviewed and returned within three weeks of submission.
 - Remote Access Users: Output should be received within a few hours. If it takes longer, please contact the Remote Access Team.
2. Output MUST match the research questions/output suggested in the proposal.
3. We recommend that output emphasize model output rather than tabular output.
4. We recommend that researchers avoid intermediate output (output that will not appear in publication). This output can certainly be created, but we discourage its release.
 - Examples of intermediate output include: tables of preliminary descriptive statistics, large numbers early regression models
 - Reasons intermediate output increase disclosure risk: similar tables based on different sub-samples may cause complementary disclosure problems because comparison of tables could reveal information about the sample and individual characteristics.
5. Review Your Output First:
 - If there is any output that you feel could lead to the identification of an individual or institution, remove it. If you have questions, discuss this with your RDC Analyst.
 - Printing individual cases is not permitted. Remove any individual level data from your output.
 - Data sets will not be released.
 - All cells with a frequency less than 5 should be removed before submitting the tables to your RDC Analyst. If possible, consider re-categorizing the variables instead of removing the observations completely.
 - Extreme values or values representing an individual must be removed. Examples include minima, maxima, medians, and modes. If a procedure, such as Proc Univariate creates extreme observations, 0, 1, 99, and 100 percentiles, those extreme values must also be removed.
 - Any procedure that produces output on an individual or institution must be removed.
6. NCHS and Census RDC Users: Output will be released in the default format produced by the statistical package (e.g., .lst files for SAS). Or output can be released as RTF files. Before submitting output for review, you must label each

NCHS RDC Disclosure Manual: Rules for Researchers

output with the following information (this can be done using a title statement or typed into an RTF):

- What is this output? (This can be a title: a regression of...)
- Who is the (sub)sample in this analysis? (black males age 20-29)
- How will output be used? (publication, presentation, etc.)

Output Review Policies

Disclosure Analysis for Tabular Output (please limit tables): Tables of summary statistics (means, variances, and related statistics) intended to describe the distributions of variables present certain special issues:

- Disclosure Analysis of Tabular Output: When cell sizes are too small (<5) in frequency tables, there are two methods for preventing disclosure:
 - a. Categorizing Variables: Researchers can combine small cells into larger categories, eliminating the particular cells that produce disclosures.
 - b. Complementary Suppression: Otherwise, the small cells will be suppressed. In order to prevent determination of the unacceptable cell figure through subtraction, other cells will also be suppressed. ANDRE will automatically perform this operation.
 - c. We require that on site researchers examine their output for small cells before asking for it to be reviewed by an RDC Analyst.
- Disclosure Analysis for Model Output: For model output, disclosure is usually not a problem if the sample contains a sufficient number of observations—establishments, households, or individuals—to start. Model output is strongly encouraged over tabular results.
- If after disclosure review, the RDC has denied release of output, you have the choice to ask that the Data Review Board (DRB) consider the issue. Their decisions are final. You will have to write a one page justification explaining why the output should be released.